

UNIVERSITÉ DE GENÈVE
Laboratoire de statistique mathématique

FACULTÉ DES SCIENCES
Professeur A. Linder

Critères
de sélection pour le choix des variables
dans l'analyse de régression

THÈSE

présentée à la faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur en statistique mathématique

par

HEINZ LINHART
d'Allemagne

THÈSE N° 1283

BERNE IMPRIMERIE STÆMPFLI & CIE
1958

La faculté des sciences, sur le préavis de Monsieur le professeur A. LINDER autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 6 mars 1958.

Le doyen: FERNAND CHODAT

THÈSE N° 1283

Extrait de la «Revue suisse de Statistique et d'Economie politique», 94^e année, 1958

UNIVERSITÉ DE GENÈVE
Laboratoire de statistique mathématique

FACULTÉ DES SCIENCES
Professeur A. Linder

Critères
de sélection pour le choix des variables
dans l'analyse de régression

THÈSE

présentée à la faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur en statistique mathématique

par

HEINZ LINHART
d'Allemagne



THÈSE N° 1283

BERNE IMPRIMERIE STÄMPFLI & CIE

1958

Critère de sélection pour le choix des variables dans l'analyse de régression

Par H. Linhart, Johannesburg

Laboratoire de statistique mathématique, Université de Genève; National Institute for Personnel Research, South African Council for Scientific and Industrial Research

1^o Aperçu historique

L'origine de la théorie de régression remonte à *Legendre* (1806) et à *Gauss* (1809) qui introduisirent le principe des moindres carrés et donnèrent ainsi une estimation des paramètres dans une analyse de régression. Les problèmes de répartition, dans le cas le plus important de la régression linéaire avec des répartitions supposées normales, furent résolus au cours des premières décennies de ce siècle, période de rapide développement de la statistique mathématique moderne. Les publications les plus importantes sont celles de *Yule* (1907), qui introduisit une nouvelle notation et discuta quelques répartitions de grands échantillons; de *Fisher* qui découvrit la répartition du coefficient de corrélation (1915), du coefficient de régression (1922), du coefficient de corrélation partielle (1924) et multiple (1928); et de *Bartlett* (1933), qui dériva, une fois de plus, les répartitions les plus importantes et fit ressortir la nécessité des hypothèses de normalité.

Le travail de *Bartlett* peut être considéré comme final, mais des généralisations et des modifications en tout sens donnèrent un nouvel essor à la statistique mathématique et conduisirent aux analyses de variance, de covariance et des composantes de variance ainsi qu'à l'analyse de la variance à plusieurs variables. Tandis que le mérite d'avoir commencé l'étude des premiers revient à *Fisher*, *Wilks* fut le premier à avoir généralisé la notion de variance au cas de plusieurs variables (1932). La répartition simultanée des variances et des covariances de répartitions normales à plusieurs variables, qui a été introduite par *Wishart* (1928), joue un rôle prépondérant dans ce domaine.

Les méthodes de *Gauss* visaient à l'évaluation de mesures entachées d'erreurs. Elles sont encore utilisées, dans ce sens, en sciences naturelles. Une relation fonctionnelle entre une variable dépendante et quelques variables indépendantes est connue ou supposée connue à l'exception de certains paramètres. Les mesures des variables ne sont pas précises; on se sert de la méthode des moindres carrés pour obtenir une estimation de ces paramètres.

En s'inspirant des méthodes en usage en sciences naturelles, des savants essayèrent de découvrir les lois qui régissent les sciences économiques et sociales. Ils firent des observations et des expériences – c'est-à-dire qu'ils firent des mensurations – et ils furent tout naturellement amenés à se servir des techniques de

régression. Ils se trouvèrent devant une situation beaucoup plus embrouillée : leurs variables avaient des variances assez considérables et dépendaient entre elles d'une façon très compliquée. Dans la plupart des cas, il n'y avait pas moyen de découvrir la véritable équation de régression ou d'augmenter la précision des résultats par des mesures plus exactes. C'est peut-être la raison pour laquelle on introduisit le nouveau terme de «prédiction», caractérisant une situation où l'accent n'était plus mis sur la découverte de lois de validité générale, mais plus modestement sur l'utilisation, pour des prédictions, des connaissances sur les interdépendances observées, et ceci en général à l'intérieur d'un champ d'investigation très limité.

La situation qui vient d'être décrite est aussi responsable de l'essai qui a été fait pour améliorer les prédictions, non en augmentant la précision des mesures, mais en faisant entrer toujours davantage de variables dans une analyse. Cependant, le résultat des applications et des expériences a montré que cette tentative n'est pas recommandable, car la prédiction est parfois moins bonne lorsqu'on inclut davantage de variables dans une analyse. Pour ne citer qu'un résultat extrême : *Guttman* (1955, p. 307) fit une analyse de régression utilisant 63 variables de prédiction dans un échantillon de 163 cas, obtenant un coefficient de corrélation multiple de 0,73. En se servant des coefficients de régression pour prédire la valeur de la variable dépendante dans un autre échantillon du même ensemble de base, il obtint – entre les valeurs observées et les valeurs prédites – un coefficient de corrélation de 0,04. Une pondération arbitraire dans la régression conduisit à un coefficient de corrélation d'environ 0,25.

Le seul procédé statistique très souvent utilisé, qui puisse, à notre connaissance, permettre de décider si une variable doit être incluse ou non dans une analyse, est celui du test de signification des coefficients de régression ; les variables correspondant à des coefficients non-significatifs sont exclues. (Cf. p.ex. *Johnson*, 1949.) Ce procédé définit implicitement la variance résiduelle comme mesure de l'efficacité du procédé de prédiction. On peut démontrer ceci : si le coefficient de régression d'une variable est nul, la variance résiduelle reste la même lorsque cette variable est incluse dans l'analyse – et vice versa). Cette définition n'est pas celle à laquelle on s'attend intuitivement. Il est clair que la puissance du procédé de prédiction – au sens habituel – pourrait décroître considérablement, mais on peut prouver que la variance résiduelle ne peut jamais augmenter si l'on inclut davantage de variables, que dès lors la puissance de la prédiction – au sens de la définition qui précède – ne peut pas diminuer.

La raison de cette incohérence réside dans le fait que la définition précitée n'est basée que sur les paramètres de l'ensemble de base et ne tient pas compte du procédé de prédiction lui-même, qui comprend l'échantillonnage et l'estimation. Elle est raisonnable si les paramètres de l'ensemble de base sont connus, mais elle devient douteuse si les paramètres doivent être estimés.

Dans cette thèse, nous proposons une nouvelle définition de la puissance du procédé de prédiction qui est basée sur les paramètres de l'ensemble de base et du procédé d'estimation. Cette définition n'a pas les défauts que nous venons de mentionner et conduit à un nouveau critère sur lequel on peut se baser pour

décider si certaines variables devraient être incluses dans une analyse de régression.

Les résultats obtenus montreront que, en dépit de ce qu'il part d'un point de vue très différent, le procédé statistique finalement suggéré sera très similaire à celui qui a été mentionné plus haut. Jusqu'à présent, on a toujours testé si le rapport de la variance résiduelle obtenue en excluant la variable, à la variance résiduelle obtenue lorsqu'on l'inclut, était égal à 1. Nous suggérons maintenant de tester si ce rapport est plus petit ou égal à une constante, elle-même supérieure à 1, et dont on peut indiquer la valeur numérique.

Cette similitude n'est que formelle. D'une façon générale, le nouveau critère est plus strict et révèle le fait qu'il est important de savoir si l'on prédit la moyenne d'une variable ou sa valeur même. Si l'on prédit la moyenne, c'est-à-dire la valeur située sur l'hyperplan de régression, le nouveau critère s'avère totalement différent du critère conventionnel: dans la plupart des cas, on n'inclura qu'une seule variable. Si l'on prédit la variable elle-même, la différence entre les deux méthodes apparaît clairement pour de petits échantillons et diminue pour de grands échantillons.

Par la suite, on verra qu'avec cette nouvelle définition, la puissance du procédé de prédiction peut diminuer lorsqu'on fait intervenir plusieurs variables, et que le défaut de la première définition est ainsi supprimé.

Un autre avantage du nouveau procédé consiste en ce qu'il ne teste pas une «hypothèse de point», contrairement à l'ancien procédé. La probabilité qu'une hypothèse de point se révèle juste, est nulle si la répartition a priori des paramètres est continue, quelle que soit la forme de cette répartition. C'est donc uniquement la grandeur de l'échantillon qui décide si l'on veut accepter ou rejeter l'hypothèse.

2° Le critère

Nous allons proposer un critère d'après lequel on peut décider si, raisonnablement, une variable (ou un groupe de r variables) doit être incluse ou exclue d'une analyse de régression donnée. Le critère est limité au cas de la théorie normale, et il est basé sur l'intervalle de confiance de la variable à prédire ou celui de sa moyenne.

On sait qu'en faisant les suppositions habituelles, on peut obtenir des limites de confiance pour x_0 et pour $E(x_0)$, la moyenne de x_0 , si l'on connaît la série des valeurs de $x_1, \dots, x_k$ et si l'on a fait, auparavant, une analyse de régression de x_0 en $x_1, \dots, x_k$.

Une supposition que l'on fait d'ordinaire pour le calcul des intervalles de confiance, veut que $x_1, \dots, x_k$ soient des valeurs fixes et non des variables aléatoires. On a cependant montré (Crow, 1955) que ces limites de confiance sont valables même si les $x_1, \dots, x_k$ sont des variables aléatoires.

Dans beaucoup d'applications, $x_1, \dots, x_k$ sont des variables aléatoires; dans les applications psychologiques, elles le sont toujours. Supposons donc que $x_0, x_1, \dots, x_k$ soient des variables aléatoires et qu'elles suivent une répartition normale à $(k+1)$ dimensions.

Si l'on se trouve en face d'une répartition normale à $(k + 1)$ dimensions et qu'on veut prédire x_0 , ou $E(x_0)$, au moyen des autres variables, on choisit un échantillon d'effectif n des $x_0, x_1, \dots, x_k$ (appelé par la suite «échantillon de régression»), et on calcule les coefficients de régression. Ensuite, on prélève quelques séries de valeurs de $x_1, x_2, \dots, x_k$ (appelées plus tard «séries de prédiction») et désignées par $y_1, y_2, \dots, y_k$, et on prédit x_0 , ou $E(x_0)$. On obtient une idée de la qualité de la prédiction en calculant l'intervalle de confiance relatif à x_0 , ou à $E(x_0)$.

Il semble alors indiqué d'utiliser le groupe de variables (parmi $x_1, \dots, x_k$) qui donnera, à la longue, les intervalles de confiance les plus étroits. Plus précisément: il faudrait choisir le groupe de variables de façon que l'espérance mathématique, $E(l)$, de la longueur de l'intervalle de confiance soit la plus petite, l'espérance devant être calculée à partir de tous les «échantillons de régression» et toutes les «séries de prédiction». Telle est l'idée qui est à la base du critère que nous proposons.

Il s'ensuit qu'il faudra discerner les cas où x_0 est l'objet de prédiction de ceux où c'est $E(x_0)$, puisque $E(l)$ prend des valeurs différentes selon l'un ou l'autre des cas.

S'il subsiste un doute quant à la décision d'inclure ou non une variable, il faut trouver le rapport de la valeur de $E(l)$, lorsque la variable est incluse, à la valeur de $E(l)$, lorsque la variable est exclue; la variable doit être incluse si ce rapport est plus petit que 1, et dans ce cas seulement. La valeur de ce rapport n'est jamais connue, car elle dépend des paramètres de l'ensemble de base. C'est pourquoi il faut tester statistiquement l'hypothèse affirmant que le rapport est plus grand que 1. La nouvelle variable doit être incluse, si, et seulement si cette hypothèse peut être rejetée.

Cela implique qu'on inclura faussement des variables dans peu de cas seulement, mais qu'on en exclura faussement assez souvent. Si ce n'est pas ce qu'on désire, on pourrait songer à choisir une erreur- α de 50%. Cela reviendrait à choisir la médiane de la répartition de la statistique utilisée (dans le cas où le rapport est égal à 1) comme point critique définissant la région de rejet. C'est un procédé optimum à appliquer dans le cas où l'on est forcé de prendre une décision et lorsqu'on pondère de manière égale les deux erreurs possibles. On pourrait le recommander dans le cas où l'échantillon utilisé est trop petit pour assurer une puissance suffisante du test.

Ce procédé peut naturellement être modifié de différentes façons. On pourrait par exemple inclure une nouvelle variable seulement si elle réduit $E(l)$ de 10%. On devrait alors tester si le rapport susmentionné est plus grand que 0,9 ou non.

Le même procédé peut être utilisé quand on doit décider si un groupe de r variables choisies parmi les k variables primitives doit être exclu ou non.

On verra que $E(l)$ est fonction de l'effectif de «l'échantillon de régression», du nombre de variables utilisées, du coefficient de confiance choisi et de la déviation standard résiduelle de l'ensemble de base. Il peut très bien augmenter si l'on inclut une variable supplémentaire, même si le coefficient de régression correspondant diffère de zéro.

3^o Notation

$\mathbf{x}$	vecteur de composantes $x_0, x_1, \dots, x_k$.
x_0	variable à prédire.
$x_1, x_2, \dots, x_k$	variables de prédiction.
$\bar{\mathbf{x}}$	vecteur de composantes $\bar{x}_0, \bar{x}_1, \dots, \bar{x}_k$.
$\bar{x}_i$ ($i = 0, 1, \dots, k$)	moyenne de x_i dans un échantillon d'effectif n (échantillon de régression).
$\mathbf{y}$	vecteur de composantes $y_1, y_2, \dots, y_k$.
$y_1, y_2, \dots, y_k$	un ensemble de valeurs de $x_1, x_2, \dots, x_k$ (série de prédiction).
$\mathbf{L}$	matrice régulière symétrique d'ordre $(k+1)$ à éléments l_{ij} .
$l_{ij} = \frac{1}{n} \sum (x_i - \bar{x}_i)(x_j - \bar{x}_j)$	où la sommation $\sum$ s'étend sur tout l'échantillon de régression.
$\mathbf{\Lambda}$	«matrice des moments» régulière symétrique d'ordre $k+1$ à éléments λ_{ij} .
λ_{ij} ($i, j = 0, 1, \dots, k$)	variances et covariances dans l'ensemble de base des $x_0, x_1, \dots, x_k$.
$\mathbf{A}_{\alpha\beta\gamma}$	matrice obtenue à partir de $\mathbf{A}$ en supprimant les lignes et colonnes contenant des éléments dont l'un des indices au moins est α, β, γ .
$\mathbf{a}_{\alpha\beta\gamma}$	vecteur obtenu à partir de $\mathbf{a}$ en omettant toutes les composantes d'indice α, β, γ .
R	coefficient de corrélation multiple.
$t_{\alpha}^{(v)}$	limite (bilatérale) de signification de la répartition de t de «Student» avec v degrés de liberté pour un seul de 100 α .
$[r]$ comme indice	signifie partout: dans le cas où r variables ont été omises.
p. ex. $R_{[r]}$	coefficient de corrélation multiple calculé lorsqu'on a omis r variables.

Les éléments de matrices inverses sont désignés par des indices supérieures. Par exemple: λ^{ij} sont les éléments de $\mathbf{\Lambda}^{-1}$; on écrira Λ_0^{-1} au lieu de $(\Lambda_0)^{-1}$. $L, L_0, A, \dots$ sont les déterminants de $\mathbf{L}_0, \mathbf{L}, \mathbf{A}, \dots$

Si a_{ij} , ($i, j = 1, 2, \dots, k$), sont les éléments de $\mathbf{A}$, les éléments de $\mathbf{A}_{\alpha\beta\gamma}^{-1}$ sont désignés par $a_{\alpha\beta\gamma}^{ij}$

($i, j = 1, 2, \dots, \alpha-1, \alpha+1, \dots, \beta-1, \beta+1, \dots, \gamma-1, \gamma+1, \dots k$).

Par exemple :

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & \cdot & \cdot & \cdot \\ a_{31} & \cdot & \cdot & \cdot \\ a_{41} & \cdot & \cdot & a_{44} \end{pmatrix} \quad \mathbf{A}_3^{-1} = \begin{pmatrix} a_3^{11} & a_3^{12} & a_3^{14} \\ a_3^{21} & a_3^{22} & a_3^{24} \\ a_3^{41} & a_3^{42} & a_3^{44} \end{pmatrix}$$

4° Espérance mathématique de la longueur des intervalles de confiance

4.1. Espérance mathématique de la longueur de l'intervalle de confiance relatif à $E(x_0)$

La longueur de l'intervalle peut être écrite sous la forme :

$$l_m = 2 t_{\alpha}^{(n-k-1)} (n-k-1)^{-\frac{1}{2}} \left(\frac{L}{L_0} \right)^{\frac{1}{2}} (1+T)^{\frac{1}{2}}. \quad (1)$$

(cf. p. ex. *Cramér*, 1946, équation 37.3.5, p. 554).

où
$$T = (\mathbf{y} - \bar{\mathbf{x}}_0)' \mathbf{L}_0^{-1} (\mathbf{y} - \bar{\mathbf{x}}_0) \quad (2)$$

désigne la distribution T généralisée de *Hotelling*.

L'espérance mathématique de l sera tout d'abord calculée pour tous les échantillons de régression, puis pour toutes les séries de prédiction :

$$E(l_m) = E_y [E_x(l_m)] \quad (3)$$

La notation est évidente.

On suppose que les variables $x_0, x_1, \dots, x_k$ ont une répartition normale régulière à $k+1$ dimensions avec, comme densité de probabilité :

$$f(x_0, x_1, \dots, x_k) = (2\pi)^{-\frac{k+1}{2}} \Lambda^{-\frac{1}{2}} \exp. \left\{ -\frac{1}{2} \mathbf{x}' \Lambda^{-1} \mathbf{x} \right\}. \quad (4)$$

Les variances et covariances l_{ij} sont alors réparties indépendamment de la moyenne $\bar{x}_i$; *Wishart* (1928) a indiqué la répartition:

$$f(l_{00}, l_{01}, \dots, l_{kk}) = C_{k+1n} \left(\frac{n^{k+1}}{2^{k+1} \Lambda} \right)^{\frac{n-1}{2}} L^{\frac{n-(k+1)-2}{2}} \exp. \left\{ -\frac{n}{2} \text{tr} \Lambda^{-1} L \right\} \quad (5)$$

pour l'ensemble des points $S^{(k+1)}$, dans un espace à $\frac{1}{2}(k+1)(k+2)$ dimensions des $l_{00}, \dots, l_{kk}$ pour lequel L est définie positive;

$$f(l_{00}, l_{01}, \dots, l_{kk}) = 0$$

pour l'ensemble complémentaire de $S^{(k+1)}$.

$$\frac{1}{C_{kn}} = \pi^{\frac{k(k-1)}{4}} \prod_{i=1}^k \Gamma\left(\frac{n-i}{2}\right). \quad (6)$$

La répartition des moyennes est donnée par:

$$f(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k) = \left(\frac{n}{2\pi} \right)^{\frac{k}{2}} \Lambda_0^{-\frac{1}{2}} \exp. \left\{ -\frac{n}{2} \bar{\mathbf{x}}_0' \Lambda_0^{-1} \bar{\mathbf{x}}_0 \right\}. \quad (7)$$

On a alors:

$$E_x \left[\left(\frac{L}{L_0} \right)^{\frac{1}{2}} (1+T)^{\frac{1}{2}} \right] = \int \left(\frac{L}{L_0} \right)^{\frac{1}{2}} (1+T)^{\frac{1}{2}} f(l_{00}, \dots, l_{kk}) f(\bar{x}_1, \dots, \bar{x}_k) \quad (8)$$

$$dl_{00} dl_{01} \dots dl_{kk} d\bar{x}_1 \dots d\bar{x}_k$$

où $\int$ remplace des intégrales multiples, et où l'intégrale, par rapport à $l_{00}, \dots, l_{kk}$ doit être étendue à l'ensemble $S^{(k+1)}$. L'intégration par rapport à $\bar{x}_1, \dots, \bar{x}_k$ va de $-\infty$ à $+\infty$.

On intégrera tout d'abord par rapport aux variables $l_{00}, l_{01}, \dots, l_{0k}$ qui ne sont contenues que dans L .

Wilks (1932, éq. 21, p.481) a montré que:

$$\begin{aligned} & \int L^{\frac{n-(k+1)-2}{2} + \frac{1}{2}} \exp. \left\{ -\frac{n}{2} \text{tr} \Lambda^{-1} L \right\} dl_{00} dl_{01} \dots dl_{0k} \\ &= \frac{C_{kn+1}}{C_{k+1n+1}} \left(\frac{2\Lambda}{n\Lambda_0} \right)^{\frac{n}{2}} L_0^{\frac{n-k-2}{2} + \frac{1}{2}} \exp. \left\{ -\frac{n}{2} \text{tr} \Lambda_0^{-1} L_0 \right\}. \end{aligned} \quad (9)$$

(L'intégrale est étendue à l'espace où L est positif, L_0 définie positive.)

En utilisant ce résultat, on obtient

$$E_x \left[\left(\frac{L}{L_0} \right)^{\frac{1}{2}} (1+T)^{\frac{1}{2}} \right] = \frac{\Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)} \left(\frac{2\Lambda}{n\Lambda_0} \right)^{\frac{1}{2}} E_x \left[(1+T)^{\frac{1}{2}} \right]. \quad (10)$$

Hsu (1938) a montré que la répartition de fréquences de T est la répartition non centrée de F :

$$f(T) = \exp. \{-\lambda\} \sum_{v=0}^{\infty} \frac{\lambda^v T^{\frac{k}{2}+v-1}}{v! B\left(\frac{k}{2} + v, \frac{n-k}{2}\right)} (1+T)^{-\frac{n}{2}-v} \quad (11)$$

où
$$\lambda = \frac{n}{2} \mathbf{Y}' \mathbf{\Lambda}_0^{-1} \mathbf{Y} \quad (12)$$

(cf. Hsu (1938) p.235, éq.12).

Puis:

$$\begin{aligned} E_x \left[(1+T)^{\frac{1}{2}} \right] &= \int_0^{\infty} (1+T)^{\frac{1}{2}} f(T) dT \\ &= \frac{\Gamma\left(\frac{n-k-1}{2}\right) \Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-k}{2}\right) \Gamma\left(\frac{n-1}{2}\right)} F\left(-\frac{1}{2}; \frac{n-1}{2}; -\lambda\right) \end{aligned} \quad (13)$$

où $F(a; b; z)$ est la série confluyente hypergéométrique.

Nous pouvons maintenant trouver $E_y[E_x(l_m)]$. Les variables auxquelles on applique l'opération sont donc y_i ($i = 1, 2, \dots, k$) qui ont une répartition normale à k dimensions avec une matrice des moments $\mathbf{\Lambda}_0$ et un vecteur moyen $\mathbf{0}$.

On sait que $\eta = \mathbf{Y}' \mathbf{\Lambda}_0^{-1} \mathbf{Y} = \frac{2\lambda}{n}$ est réparti selon χ^2 avec k degrés de liberté. (Cf. p. ex. Cramér (1946) p. 319, ex.15.)

L'espérance mathématique de $F\left(-\frac{1}{2}; \frac{n-1}{2}; -\frac{n}{2}\eta\right)$ est dès lors obtenue en mettant le v -ème moment de χ^2 , à savoir $k(k+2) \dots (k+2v-2)$ à la place de η^v dans la série confluyente hypergéométrique.

$$\begin{aligned} E_y \left[F\left(-\frac{1}{2}; \frac{n-1}{2}; -\frac{n}{2}\eta\right) \right] &= \int_{-\infty}^{+\infty} F\left(-\frac{1}{2}; \frac{n-1}{2}; -\frac{n}{2}\eta\right) f(y_1, \dots, y_k) dy_1 \dots dy_k \\ &= (n+1)^{-\frac{k}{2}} F\left(\frac{k}{2}, \frac{n}{2}; \frac{n-1}{2}; \frac{n}{n+1}\right) \end{aligned} \quad (14)$$

où $F(a, b; c; z)$ est la série hypergéométrique.

En réunissant les termes constants, on a ainsi:

$$E(l_m) = \frac{2 t_{\alpha}^{(n-k-1)} \Gamma\left(\frac{n}{2}\right)}{(n-k-1)^{\frac{1}{2}} (n+1)^{\frac{k}{2}} \Gamma\left(\frac{n-1}{2}\right)} F\left(\frac{k}{2}, \frac{n}{2}; \frac{n-1}{2}; \frac{n}{n+1}\right) \left(\frac{2A}{nA_0}\right)^{\frac{1}{2}} \quad (15)$$

4.2. Espérance mathématique de la longueur de l'intervalle de confiance relatif à x_0

La longueur de l'intervalle est maintenant:

$$l_v = 2 t_{\alpha}^{(n-k-1)} (n-k-1)^{-\frac{1}{2}} \left(\frac{L}{L_0}\right)^{\frac{1}{2}} (n+1+T)^{\frac{1}{2}}. \quad (16)$$

On obtient comme dans 4.1.:

$$\begin{aligned} E\left[\left(\frac{L}{L_0}\right)^{\frac{1}{2}} (n+1+T)^{\frac{1}{2}}\right] &= \\ &= \frac{\Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)} \left(\frac{2A}{nA_0}\right)^{\frac{1}{2}} E\left[(n+1+T)^{\frac{1}{2}}\right]. \end{aligned} \quad (17)$$

On peut prouver que:

$$E_y[\lambda^v \exp. \{-\lambda\}] = \frac{\Gamma\left(\frac{k}{2} + v\right)}{\Gamma\left(\frac{k}{2}\right)} \left(\frac{n}{n+1}\right)^v (n+1)^{-\frac{k}{2}}. \quad (18)$$

C'est pourquoi:

$$\begin{aligned} E\left[(n+1+T)^{\frac{1}{2}}\right] &= \\ &= \frac{1}{\Gamma\left(\frac{n-k}{2}\right) \Gamma\left(\frac{k}{2}\right) (n+1)^{\frac{k}{2}}} \int_0^{\infty} T^{\frac{k}{2}-1} (n+1+T)^{\frac{1}{2}} (1+T)^{-\frac{n}{2}} \\ &\quad \sum_{v=0}^{\infty} \left(\frac{n}{n+1}\right)^v \left(\frac{T}{1+T}\right)^v \frac{1}{v!} \Gamma\left(\frac{n}{2} + v\right) dT \end{aligned}$$

$$= \frac{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{n-k-1}{2}\right) (n+1)^{\frac{1}{2}}}{\Gamma\left(\frac{n-1}{2}\right) \Gamma\left(\frac{n-k}{2}\right)} \quad (19)$$

et

$$E(l_v) = \frac{2 t_{\alpha}^{(n-k-1)} (n+1)^{\frac{1}{2}} \Gamma\left(\frac{n}{2}\right)}{(n-k-1)^{\frac{1}{2}} \Gamma\left(\frac{n-1}{2}\right)} \left(\frac{2A}{nA_0}\right)^{\frac{1}{2}}. \quad (20)$$

4.3. Estimations non biaisées de $E(l)$

Wilks (1932) a montré que

$$E\left[\left(\frac{L}{L_0}\right)^{\frac{1}{2}}\right] = \frac{\Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)} \left(\frac{2A}{nA_0}\right)^{\frac{1}{2}}. \quad (21)$$

Les estimations non biaisées de $E(l)$ sont par conséquent, dans le cas d'intervalles de confiance relatifs à $E(x_0)$:

$$\hat{E}(l_m) = \frac{2 t_{\alpha}^{(n-k-1)} \Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{n-k-1}{2}\right) F\left(\frac{k}{2}, \frac{n}{2}; \frac{n-1}{2}; \frac{n}{n+1}\right)}{(n-k-1)^{\frac{1}{2}} (n+1)^{\frac{k}{2}} \Gamma\left(\frac{n-1}{2}\right) \Gamma\left(\frac{n-k}{2}\right)} \left(\frac{L}{L_0}\right)^{\frac{1}{2}} \quad (22)$$

dans les cas d'intervalles de confiance relatifs à x_0 on a:

$$\hat{E}(l_v) = \frac{2 t_{\alpha}^{(n-k-1)} (n+1)^{\frac{1}{2}} \Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{n-k-1}{2}\right)}{(n-k-1)^{\frac{1}{2}} \Gamma\left(\frac{n-1}{2}\right) \Gamma\left(\frac{n-k}{2}\right)} \left(\frac{L}{L_0}\right)^{\frac{1}{2}}. \quad (23)$$

5^o Répartition de l'estimation d'une caractéristique

5.1. Une caractéristique

Dans le chapitre 2, on a établi qu'il faudrait partir de la valeur $\frac{E(l)}{E_{[r]}(l)}$ pour décider si r parmi les k variables de prédiction devraient être exclues ou non. En plus des constantes connues, ce rapport contient toujours la caractéristique

$$\left(\frac{A_{0\ k+1-r\dots k}}{A_{k+1-r\dots k}} \right)^{\frac{1}{2}} = \gamma_{[r]}^{\frac{1}{2}}. \quad (24)$$

Il est donc nécessaire de tester des hypothèses concernant $\gamma_{[r]}$. Une estimation de la caractéristique qui, intuitivement, semble appelée à être utilisée pour ce genre de tests est

$$\frac{L_{0\ k+1-r\dots k}}{L_{k+1-r\dots k}} = c_{[r]}. \quad (25)$$

Dans le chapitre suivant, on établira la répartition de $c_{[r]}$.

5.2. Preuve d'un lemme

Le lemme suivant sera démontré en premier lieu.

Soit $\mathbf{A}$ une matrice symétrique définie positive d'ordre k , avec les éléments a_{ij} , ($i, j = 1, 2, \dots, k$), et soit $\mathbf{u}$ un vecteur de composantes u_i , ($i = 1, 2, \dots, k$).

Alors, les k différentes formes quadratiques

$$\mathbf{u}'_{12\dots t} \mathbf{A}_{12\dots t}^{-1} \mathbf{u}_{12\dots t} \quad (26)$$

(où l'ensemble des indices $1, 2, \dots, t$ peut être vide sinon $t = 1, 2, \dots, k-1$) peuvent être transformées simultanément en sommes de carrés.

La transformation ayant la propriété requise est

$$\mathbf{u} = \mathbf{A} \mathbf{B} \mathbf{v} \quad (27)$$

où $\mathbf{B}$ est une matrice triangulaire indiquée plus bas, et

$$\mathbf{B} = \mathbf{A}^{-\frac{1}{2}} \quad (28)$$

Pour établir la preuve, on a besoin d'une identité concernant des mineurs, due à *Stouffer* (1932), qui peut être généralisée et écrite sous la forme:

$$a^{ij} = a_i^{ij} + \frac{a^{li} a^{lj}}{a^{ll}} \quad (29)$$

($l \neq i, j$; $l, i, j = 1, 2, \dots, k$).

Par application répétée de cette formule, on trouve:

$$\mathbf{A}_{12\dots t}^{-1} = \mathbf{B}_{12\dots t} \mathbf{B}'_{12\dots t} \quad (30)$$

$$\text{où } \mathbf{B} = \begin{pmatrix} \frac{a^{11}}{\sqrt{a^{11}}} & 0 & 0 & \dots & 0 \\ \frac{a^{12}}{\sqrt{a^{11}}} & \frac{a_1^{22}}{\sqrt{a_1^{22}}} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{a^{1k}}{\sqrt{a^{11}}} & \frac{a_1^{2k}}{\sqrt{a_1^{22}}} & \vdots & \vdots & \frac{a_{12\dots k-1}^{kk}}{\sqrt{a_{12\dots k-1}^{kk}}} \end{pmatrix} \quad (31)$$

La transformation:

$$\mathbf{u} = \mathbf{A} \mathbf{B} \mathbf{v} \quad (32)$$

donne:

$$\mathbf{u}'_{12\dots t} \mathbf{A}_{12\dots t}^{-1} \mathbf{u}_{12\dots t} = (\mathbf{A} \mathbf{B} \mathbf{v})'_{12\dots t} \mathbf{A}_{12\dots t}^{-1} (\mathbf{A} \mathbf{B} \mathbf{v})_{12\dots t}. \quad (33)$$

On peut vérifier que

$$(\mathbf{v}' \mathbf{B}^{-1})_{12\dots t} = \mathbf{v}'_{12\dots t} \mathbf{B}_{12\dots t}^{-1}. \quad (34)$$

Par conséquent,

$$\mathbf{u}'_{12\dots t} \mathbf{A}_{12\dots t}^{-1} \mathbf{u}_{12\dots t} = \mathbf{v}'_{12\dots t} \mathbf{B}_{12\dots t}^{-1} \mathbf{A}_{12\dots t}^{-1} \mathbf{B}_{12\dots t}^{-1'} \mathbf{v}_{12\dots t} = \mathbf{v}'_{12\dots t} \mathbf{v}_{12\dots t}. \quad (35)$$

Ceci, avec

$$\mathbf{A}^{-1} = \mathbf{B} \mathbf{B}'$$

et par conséquent $\mathbf{A}^{-1} = \mathbf{B}^2$, établit le lemme.

5.3. Evaluation d'une intégrale

Nous allons maintenant évaluer l'intégrale multiple

$$I_{[r]} = \int \frac{L^{\frac{n+2h-(k+1)-2}{2}}}{L_0^h} \exp. \left\{ -\frac{n}{2} \text{tr} \mathbf{A}^{-1} \mathbf{L} \right\} \frac{dl_{0\ k+1-r} dl_{1\ k+1-r} \dots dl_{k+1-r\ k+1-r}}{dl_{0\ k+1-r+1} \dots dl_{kk}} \quad (36)$$

Cette intégration est telle que $L, L_k, \dots, L_{k+1-r+1\dots k}$ sont positifs pour des $L_{k+1-r\dots k}$ définies positives.

Nous allons d'abord intégrer $I_{[1]}$, le résultat pour $I_{[r]}$ sera obtenu en raisonnant par induction.

$$\begin{aligned}
 I_{[1]} &= \int \frac{L^{\frac{n+2h-(k+1)-2}{2}}}{L_0^h} \exp. \left\{ -\frac{n}{2} \text{tr} \Lambda^{-1} \mathbf{L} \right\} dl_{0k} dl_{1k} \dots dl_{kk} \\
 &= \exp. \left\{ -\frac{n}{2} \text{tr} (\Lambda^{-1})_k \mathbf{L}_k \right\} \frac{L_k^{\frac{n+2h-(k+1)-2}{2}}}{L_{0k}^h} \int \frac{(l_{kk} - \mathbf{l}' \mathbf{L}_k^{-1} \mathbf{l})^{\frac{n+2h-(k+1)-2}{2}}}{(l_{kk} - \mathbf{l}'_0 \mathbf{L}_{0k}^{-1} \mathbf{l}_0)^h} \\
 &\quad \exp. \left\{ -\frac{n}{2} \lambda^{kk} l_{kk} - n \boldsymbol{\lambda}' \mathbf{l} \right\} dl_{0k} \dots dl_{kk}.
 \end{aligned} \tag{37}$$

($\mathbf{l}$ a les composantes $l_i \equiv l_{ik}$, $i = 0, 1, 2, \dots, k-1$).

$$\boldsymbol{\lambda}' = \{ \lambda^{0k}, \lambda^{1k}, \dots, \lambda^{k-1k} \} \tag{38}$$

En utilisant le lemme de 5.2, les deux formes quadratiques sous le signe d'intégration peuvent être transformées simultanément en sommes de carrés par

$$\mathbf{l} = \mathbf{L}_k \mathbf{H} \mathbf{z} \tag{39}$$

le Jacobien étant

$$L_k^{\frac{1}{2}}$$

$$\mathbf{z}' = \{ z_0, z_1, \dots, z_{k-1} \}. \tag{40}$$

$\mathbf{H}$ correspond à la matrice $\mathbf{B}$ donnée dans 5.2. comme formule (31)

$$\begin{aligned}
 I_{[1]} &= \exp. \left\{ -\frac{n}{2} \text{tr} (\Lambda^{-1})_k \mathbf{L}_k \right\} \frac{L_k^{\frac{n+2h-k-2}{2}}}{L_{0k}^h} \\
 &\quad \int \frac{(l_{kk} - \mathbf{z}' \mathbf{z})^{\frac{n+2h-(k+1)-2}{2}}}{(l_{kk} - \mathbf{z}'_0 \mathbf{z}_0)^h} \exp. \left\{ -\frac{n}{2} \lambda^{kk} l_{kk} - n \boldsymbol{\lambda}' \mathbf{L}_k \mathbf{H} \mathbf{z} \right\} \\
 &\quad dl_{kk} dz_0 \dots dz_{k-1}.
 \end{aligned} \tag{41}$$

En premier lieu, on effectuera l'intégration par rapport à z_0 .

Le vecteur-ligne $n \lambda' L_k H$ sera appelé

$$\zeta' = \{\zeta_0, \zeta_1, \dots, \zeta_{k-1}\}. \quad (42)$$

On peut vérifier que

$$\zeta_0 = n \lambda^{0k} \frac{1}{\sqrt{l_k^{00}}}. \quad (43)$$

Avec les abréviations $l_{kk} - \mathbf{z}'_0 \mathbf{z}_0 = a$ et $\frac{1}{2}(n + 2h - k - 3) = \alpha$,

la première intégration est de la forme :

$$I = a^{-h} \int_{-\sqrt{a}}^{+\sqrt{a}} (a - z_0^2)^\alpha \exp. \{-\zeta_0 z_0\} dz_0. \quad (44)$$

En développant en série selon le théorème binomial et en intégrant terme à terme, on obtient :

$$\begin{aligned} I &= 2 a^{\alpha h} \left\{ a^{\frac{1}{2}} \left[\binom{\alpha}{1} - \frac{1}{3} \binom{\alpha}{1} + \frac{1}{5} \binom{\alpha}{2} - \frac{1}{7} \binom{\alpha}{3} + \dots \right. \right. \\ &\quad \left. \left. + \frac{a^{\frac{3}{2}} \zeta_0^2}{2!} \left[\frac{1}{3} \binom{\alpha}{0} - \frac{1}{5} \binom{\alpha}{1} + \frac{1}{7} \binom{\alpha}{2} - \dots + \dots \right] \right\} \\ &= a^{\alpha h} \Gamma(\alpha + 1) \left\{ a^{\frac{1}{2}} \frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{2\alpha+3}{2}\right)} + a^{\frac{3}{2}} \frac{\zeta_0^2}{2!} \frac{\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{2\alpha+5}{2}\right)} + \dots \right\} \\ &= a^{\alpha h + \frac{1}{2}} \pi^{\frac{1}{2}} \frac{\Gamma(\alpha + 1)}{\Gamma\left(\frac{2\alpha+3}{2}\right)} \left\{ 1 + \frac{a \zeta_0^2}{2} \frac{1}{2\alpha+3} + \left(\frac{a \zeta_0^2}{2}\right)^2 \frac{1}{2!} \frac{1}{(2\alpha+3)(2\alpha+5)} \right. \\ &\quad \left. + \dots \right\} \end{aligned} \quad (45)$$

On introduit ensuite la transformation :

$$w = a^{\frac{1}{2}} = (l_{kk} - \mathbf{z}'_0 \mathbf{z}_0)^{\frac{1}{2}} \quad (46)$$

où

$$2w dw = dl_{kk}.$$

$$I_{[1]} = 2\pi^{\frac{1}{2}} \frac{\Gamma\left(\frac{n+2h-(k+1)}{2}\right)}{\Gamma\left(\frac{n+2h-k}{2}\right)} \frac{L_k^{\frac{n+2h-k-2}{2}}}{L_{0k}^h} \exp. \left\{ -\frac{n}{2} \operatorname{tr} (\Lambda^{-1})_k \mathbf{L}_k \right\} \\ \cdot \int_{-\infty}^{+\infty} \exp. \left\{ -\boldsymbol{\zeta}'_0 \mathbf{z}_0 - \frac{n}{2} \lambda^{kk} \mathbf{z}'_0 \mathbf{z}_0 \right\} dz_1 \dots dz_{k-1} \quad (47)$$

$$\int_0^\infty \exp. \left\{ -\frac{n}{2} \lambda^{kk} w^2 \right\} w^{n-k-1} \left\{ 1 + \frac{w^2 \zeta_0^2}{2} \frac{1}{2\alpha+3} + \left(\frac{w^2 \zeta_0^2}{2} \right)^2 \frac{1}{2!(2\alpha+3)(2\alpha+5)} \right. \\ \left. + \dots \right\} dw.$$

Comme

$$\int_{-\infty}^{+\infty} \exp. \left\{ -\boldsymbol{\zeta}'_0 \mathbf{z}_0 - \frac{n}{2} \lambda^{kk} \mathbf{z}'_0 \mathbf{z}_0 \right\} dz_1 \dots dz_{k-1} = \left(\frac{2\pi}{n\lambda^{kk}} \right)^{\frac{k-1}{2}} \exp. \left\{ \frac{1}{2n\lambda^{kk}} \boldsymbol{\zeta}'_0 \boldsymbol{\zeta}_0 \right\} \quad (48)$$

(cf. p.ex. *Cramér*, 1946, 11.12.1.a, p.118)

$$\text{et} \quad \int_0^\infty w^\nu \exp. \left\{ -\frac{n}{2} \lambda^{kk} w^2 \right\} dw = \frac{\pi^{\frac{1}{2}} \Gamma(\nu+1)}{\Gamma\left(\frac{\nu+2}{2}\right) 2^{\nu+1}} \left(\frac{n}{2} \lambda^{kk} \right)^{-\frac{\nu+1}{2}} \quad (49)$$

(cf. p.ex. *Cramér*, 1946, 10.5.1., p. 99)

on a

$$I_{[1]} = \pi^{\frac{k+1}{2}} \frac{\Gamma\left(\frac{n+2h-(k+1)}{2}\right) \Gamma(n-k)}{\Gamma\left(\frac{n+2h-k}{2}\right) \Gamma\left(\frac{n-k+1}{2}\right)} \left(\frac{n}{2} \lambda^{kk} \right)^{-\frac{n-1}{2}} \cdot 2^{-n+k+1} \cdot \frac{L_k^{\frac{n+2h-k-2}{2}}}{L_{0k}^h} \\ \cdot F\left(\frac{n-k}{2}; \frac{n+2h-k}{2}; \frac{\zeta_0^2}{2n\lambda^{kk}}\right) \exp. \left\{ -\frac{n}{2} \operatorname{tr} (\Lambda^{-1})_k \mathbf{L}_k + \frac{1}{2n\lambda^{kk}} \boldsymbol{\zeta}'_0 \boldsymbol{\zeta}_0 \right\}. \quad (50)$$

En utilisant l'abréviation $\beta_{[1]} = -\frac{n}{2} \frac{(\lambda^{0k})^2}{\lambda^{kk}}$ le résultat peut être écrit sous la forme :

$$I_{[1]} = \pi^{\frac{k}{2}} \frac{\Gamma\left(\frac{n+2h-(k+1)}{2}\right) \Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{n+2h-k}{2}\right)} \left(\frac{n}{2} \lambda^{kk}\right)^{-\frac{n-1}{2}} \frac{L_k^{\frac{n+2h-k-2}{2}}}{L_{0k}^h} \\ F\left(h; \frac{n+2h-k}{2}; \beta_{[1]} \frac{L_k}{L_{0k}}\right) \exp. \left\{ -\frac{n}{2} \text{tr} (\Lambda^{-1})_k L_k + \frac{1}{2n \lambda^{kk}} \boldsymbol{\zeta}' \boldsymbol{\zeta} \right\}. \quad (51)$$

Puis :

$$\begin{aligned} \boldsymbol{\zeta}' \boldsymbol{\zeta} &= n^2 \boldsymbol{\lambda}' L_k H H' L_k \boldsymbol{\lambda} \\ &= n^2 \boldsymbol{\lambda}' L_k \boldsymbol{\lambda} \\ &= n^2 \text{tr} (\boldsymbol{\lambda} \boldsymbol{\lambda}' L_k) \end{aligned} \quad (52)$$

(utiliser formule (30) dans 5.2.)

et

$$\Lambda_k^{-1} = (\Lambda^{-1})_k - \frac{1}{\lambda^{kk}} \boldsymbol{\lambda} \boldsymbol{\lambda}' \quad (53)$$

(voir formule (29) dans 5.2.).

Par conséquent :

$$\begin{aligned} \exp. \left\{ -\frac{n}{2} \text{tr} (\Lambda^{-1})_k L_k + \frac{1}{2n \lambda^{kk}} \boldsymbol{\zeta}' \boldsymbol{\zeta} \right\} \\ = \exp. \left\{ -\frac{n}{2} \text{tr} \left[\left\{ (\Lambda^{-1})_k - \frac{1}{\lambda^{kk}} \boldsymbol{\lambda} \boldsymbol{\lambda}' \right\} L_k \right] \right\} \\ = \exp. \left\{ -\frac{n}{2} \text{tr} \Lambda_k^{-1} L_k \right\} \end{aligned} \quad (54)$$

et on a

$$I_{[1]} = \pi^{\frac{k}{2}} \frac{\Gamma\left(\frac{n+2h-(k+1)}{2}\right) \Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{n+2h-k}{2}\right) \left(\frac{n}{2} \lambda^{kk}\right)^{\frac{n-1}{2}}} \exp. \left\{ -\frac{n}{2} \text{tr} \Lambda_k^{-1} L_k \right\} \\ \frac{L_k^{\frac{n+2h-k-2}{2}}}{L_{0k}^h} - F\left(h; \frac{n+2h-k}{2}; \beta_{[1]} \frac{L_k}{L_{0k}}\right). \quad (55)$$

On peut raisonner par induction, en utilisant le résultat pour $I_{[1]}$, que :

$$\begin{aligned}
 I_{[r]} &= \int \frac{L^{\frac{n+2h-(k+1)-2}{2}}}{L_0^h} \exp. \left\{ -\frac{n}{2} \text{tr } \Lambda^{-1} \mathbf{L} \right\} dl_{0\ k+1-r} \dots dl_{kk} \\
 &= \frac{C_{k+1-r\ n}}{C_{k+1\ n}} \frac{\Gamma\left(\frac{n-(k+1-r)}{2}\right) \Gamma\left(\frac{n+2h-(k+1)}{2}\right)}{\Gamma\left(\frac{n-(k+1)}{2}\right) \Gamma\left(\frac{n+2h-(k+1-r)}{2}\right)} \left[\left(\frac{n}{2}\right)^r \frac{A_{k+1-r\dots k}}{A} \right]^{\frac{n-1}{2}} \\
 &\quad \cdot \exp. \left\{ -\frac{n}{2} \text{tr } \Lambda_{k+1-r\dots k}^{-1} \mathbf{L}_{k+1-r\dots k} \right\} \frac{L_{k+1-r\dots k}^{\frac{n+2h-(k+1-r)-2}{2}}}{L_{0\ k+1-r\dots k}^h} \\
 &\quad \cdot F\left(h; \frac{n+2h-(k+1-r)}{2}; \beta_{[r]} \frac{L_{k+1-r\dots k}}{L_{0\ k+1-r\dots k}}\right)
 \end{aligned} \tag{56}$$

$$\text{où} \quad \beta_{[r]} = -\frac{n}{2} \left(\frac{A_0}{A} - \frac{A_{0\ k+1-r\dots k}}{A_{k+1-r\dots k}} \right). \tag{57}$$

5.4. La répartition de $c_{[r]}$

Nous nous proposons de déterminer maintenant les moments de la répartition de $c_{[r]}$.

$$\begin{aligned}
 E(c_{[r]}^h) &= C_{k+1\ n} \left(\frac{n^{k+1}}{2^{k+1} A} \right)^{\frac{n-1}{2}} \int \int \frac{L^{\frac{n+2h-(k+1)-2}{2}}}{L_0^h} \exp. \left\{ -\frac{n}{2} \text{tr } \Lambda^{-1} \mathbf{L} \right\} \\
 &\quad dl_{0\ k+1-r} \dots dl_{kk} \cdot \frac{L_{0\ k+1-r\dots k}^h}{L_{k+1-r\dots k}^h} dl_{00} \dots dl_{k-r\ k-r}.
 \end{aligned} \tag{58}$$

Le membre droit sera tout d'abord intégré par rapport aux variables qui sont seulement contenues dans L et L_0 ; pour ce faire, on utilisera la formule (56). Ensuite, on intégrera par rapport aux variables restantes en utilisant les solutions trouvées par *Wilks* (1932). L'intégration par rapport aux premières variables est étendue au domaine où $L, L_k, \dots, L_{k+1-r+1\dots k}$ sont positifs pour $L_{k+1-r\dots k}$ définies positives, les deuxièmes intégrations sont étendues à l'espace où $L_{k+1-r\dots k}$ est définie positive. L'intégrale entière est donc prise sur l'espace $S^{(k+1)}$, comme il se doit.

$$E(c_{[r]}^h) = \left(\frac{n^{k+1}}{2^{k+1} A} \right)^{n-1} C_{k+1-r, n} \frac{\Gamma\left(\frac{n-(k+1-r)}{2}\right) \Gamma\left(\frac{n+2h-(k+1)}{2}\right)}{\Gamma\left(\frac{n-(k+1)}{2}\right) \Gamma\left(\frac{n+2h-(k+1-r)}{2}\right)}$$

$$\left[\left(\frac{n}{2} \right)^r \frac{A_{k+1-r \dots k}}{A} \right]^{-\frac{n-1}{2}} \int L_{k+1-r \dots k}^{\frac{n-(k+1-r)-2}{2}} \exp. \left\{ -\frac{n}{2} tr \Lambda_{k+1-r \dots k}^{-1} L_{k+1-r \dots k} \right\}$$

$$F\left(h; \frac{n+2h-(k+1-r)}{2}; \beta_{[r]} \frac{L_{k+1-r \dots k}}{L_{0 \ k+1-r \dots k}}\right) dl_{00} \dots dl_{k-r \ k-r}. \quad (59)$$

Wilks (1932) a évalué des intégrales ayant la forme de celles qui sont nécessaires pour l'intégration terme à terme de l'expression ci-dessus :

$$\int \frac{L_{k+1-r \dots k}^{\frac{n-(k+1-r)-2}{2} + \nu}}{L_{0 \ k+1-r \dots k}^{\nu}} \exp. \left\{ -\frac{n}{2} tr \Lambda_{k+1-r \dots k}^{-1} L_{k+1-r \dots k} \right\} dl_{00} \dots dl_{k-r \ k-r}$$

$$= \frac{1}{C_{k+1-r, n}} \left(\frac{n^{k+1-r}}{2^{k+1-r} A_{k+1-r \dots k}} \right)^{-\frac{n-1}{2}} \left(\frac{A_{k+1-r \dots k}}{A_{0 \ k+1-r \dots k}} \frac{2}{n} \right)^{\nu} \frac{\Gamma\left(\frac{n-(k+1-r)}{2} + \nu\right)}{\Gamma\left(\frac{n-(k+1-r)}{2}\right)}$$

(intégration telle que $L_{k+1-r \dots k}$ est définie positive). (60)

En se servant de cette formule, on obtient :

$$E(c_{[r]}^h) = \frac{\Gamma\left(\frac{n-(k+1-r)}{2}\right) \Gamma\left(\frac{n-(k+1)}{2} + h\right)}{\Gamma\left(\frac{n-(k+1-r)}{2} + h\right) \Gamma\left(\frac{n-(k+1)}{2}\right)} \frac{n^{-(k+1-r)}}{\gamma_{[r]}^{\frac{n-(k+1-r)}{2}}} \quad (61)$$

$$F\left(\frac{n-(k+1-r)}{2}, \frac{n-(k+1-r)}{2}; \frac{n-(k+1-r)}{2} + h; (1-\gamma_{[r]})\right)$$

où
$$\gamma_{[r]} = \frac{A A_{0 \ k+1-r \dots k}}{A_0 A_{k+1-r \dots k}}. \quad (62)$$

On peut voir que la répartition fournissant les moments obtenus a pour densité de probabilité:

$$f(c_{[r]}) = \frac{\gamma_{[r]}^{\frac{n-(k+1-r)}{2}}}{B\left(\frac{n-k-1}{2}, \frac{r}{2}\right)} c_{[r]}^{\frac{n-k-1}{2}-1} (1-c_{[r]})^{\frac{r}{2}-1}$$

$$F\left[\frac{n-(k+1-r)}{2}, \frac{n-(k+1-r)}{2}; \frac{r}{2}; (1-\gamma_{[r]})(1-c_{[r]})\right]. \quad (63)$$

La répartition obtenue est celle de $1-R^2$ où R est le coefficient de corrélation multiple. On peut s'en convaincre lorsqu'on remplace, dans la formule (A) de Fisher (1928) à la page 660, $1-R^2$ par $c_{[r]}$, $1-\rho^2$ par $\gamma_{[r]}$, n_1 par r et n_2 par $n-k-1$.

6° Application

6.1. Remarques en vue d'applications

Comme nous l'avons affirmé dans le chapitre 2, le critère sur lequel on se base pour décider si k variables de prédiction sinon un certain groupe de $k-r$ parmi elles, doivent être utilisées, est de la forme: Tester l'hypothèse $E(l) > E_{[r]}(l)$ et utiliser k variables si cette hypothèse peut être rejetée.

Dans le chapitre 4, on a montré que $E(l)$ prend des valeurs différentes selon ce qu'on doit prédire, x_0 ou $E(x_0)$. Il est, par conséquent, important de décider par avance si x_0 ou $E(x_0)$ est le centre d'intérêt, car les deux procédés respectifs seront différents.

Tester $H: E(l) > E_{[r]}(l)$ équivaut à:

(prédiction de x_0)

(prédiction de $E(x_0)$)

$$H: (1-\gamma_{[r]})^{\frac{1}{2}} < \left(1 - \frac{\varepsilon_{k-rn}}{\varepsilon_{kn}}\right)^{\frac{1}{2}} \quad H: (1-\gamma_{[r]})^{\frac{1}{2}} < \left(1 - \frac{\delta_{k-rn}}{\delta_{kn}}\right)^{\frac{1}{2}} \quad (64)$$

avec

$$\varepsilon_{kn} = \frac{n+1}{n-k-1} [t_{.05}^{(n-k-1)}]^2 \quad \text{et} \quad \delta_{kn} = \frac{F^2\left(\frac{k}{2}, \frac{n}{2}; \frac{n-1}{2}; \frac{n}{n+1}\right)}{(n-k-1)(n+1)^k} [t_{.05}^{(n-k-1)}]^2. \quad (65)$$

Conformément aux résultats de 5, la grandeur $(1 - c_{[r]})^{\frac{1}{2}} = \left(\frac{R^2 - R_{[r]}^2}{1 - R_{[r]}^2} \right)^{\frac{1}{2}}$ est répartie comme un coefficient de corrélation multiple obtenu dans un échantillon d'effectif $N = n - k + r$ tiré d'un ensemble de base avec r «variables indépendantes» ayant une corrélation multiple $(1 - \gamma_{[r]})^{\frac{1}{2}}$.

L'hypothèse mentionnée ci-dessus doit donc être rejetée si dans le graphique (relatif aux r variables indépendantes) représentant les limites de confiance inférieures pour les corrélations multiples, le point

$$\begin{array}{cc} \text{(prédiction de } x_0) & \text{(prédiction de } E(x_0)) \\ \left[\left(\frac{R^2 - R_{[r]}^2}{1 - R_{[r]}^2} \right)^{\frac{1}{2}} ; \left(1 - \frac{\varepsilon_{k-rn}}{\varepsilon_{kn}} \right)^{\frac{1}{2}} \right] & \left[\left(\frac{R^2 - R_{[r]}^2}{1 - R_{[r]}^2} \right)^{\frac{1}{2}} ; \left(1 - \frac{\delta_{k-rn}}{\delta_{kn}} \right)^{\frac{1}{2}} \right] \end{array}$$

(66)

est situé en dessous de la courbe marquée N .

Des graphiques avec limites de confiance inférieures pour des corrélations multiples sont reproduits dans l'appendice, ainsi que des tables pour ε_{kn} et δ_{kn} .

Pour obtenir une estimation non biaisée de $E(l)$, il faut calculer:

$$\begin{array}{cc} \text{(prédiction de } x_0) & \text{(prédiction de } E(x_0)) \\ \hat{E}(l_v) = [\varepsilon_{kn} l_{00} (1 - R^2)]^{\frac{1}{2}} \alpha_{kn} & \hat{E}(l_m) = [\delta_{kn} l_{00} (1 - R^2)]^{\frac{1}{2}} \alpha_{kn} \end{array} \quad (67)$$

où

$$\alpha_{kn} = 2 \frac{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{n-k-1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right) \Gamma\left(\frac{n-k}{2}\right)} \quad (68)$$

a été calculé et se trouve, sous forme de table, dans l'appendice.

6.2. Exemple d'application

Dans plusieurs mines d'or du Witwatersrand, Verwey (1957) a soumis des candidats à l'apprentissage à certains tests. Une batterie de 6 tests a été appliquée à ces candidats:

A/2	test d'intelligence générale,
A/8	test de capacité arithmétique,
A/3	} tests des connaissances scientifiques élémentaires,
A/9	
A/5	} tests du pouvoir de vision spatiale.
A/4	

Pendant la période de formation (un à cinq ans après le début de l'apprentissage) le succès des apprentis fut déterminé. On pensait qu'une évaluation combinée basée sur plusieurs tests théoriques et pratiques révélerait au mieux le succès, que des apprentis ayant obtenu de bonnes notes dans ces conditions seraient les plus capables. Le but à atteindre était d'obtenir un instrument pour la sélection des candidats les plus qualifiés. Le problème était donc de choisir quelques tests dans la batterie originale et de trouver comment pondérer des résultats des différents tests choisis afin d'obtenir les «meilleures prédictions» de succès. La technique habituelle fut utilisée, à savoir l'analyse de régression linéaire.

Les résultats des tests dans un groupe de 24 menuisiers conduisirent à la matrice de corrélation suivante:

	A/8	A/4	A/5	A/9	A/3	A/2	Succès
A/8	1	-.029618	.009318	.292254	-.038086	.311069	.475921
A/4		1	.366764	.607800	.561212	.238246	.385167
A/5			1	.320360	.419225	.374762	.262282
A/9				1	.310052	.279051	.489838
A/3					1	.423038	.178952
A/2						1	.122754
Succès							1

La variance de l'estimateur du succès était:

$$l_{00} = 242.7$$

En regardant de près cette matrice, on remarque que les variables A/9, A/8, A/4 et A/5 ont les plus fortes corrélations avec l'estimateur du succès. La corrélation est forte entre A/9 et A/4, elle est moyenne entre A/9 et A/8; A/8 et A/4 ont, cependant, une corrélation presque nulle. On peut s'attendre, par conséquent, à l'augmentation la plus grande des corrélations multiples lorsque les variables sont prises dans l'ordre:

$$A/8, A/4, A/5, A/9, \dots$$

On pourrait alors poser les questions suivantes :

est-il judicieux d'utiliser	ou est-il préférable d'utiliser seulement
1. A/8, A/4, A/5, A/9, A/3, A/2	A/8, A/4, A/5, A/9
2. A/8, A/4, A/5, A/9	A/8, A/4, A/5
3. A/8, A/4, A/5	A/8, A/4
4. A/8, A/4	A/8

On doit trouver l'inverse de la matrice de corrélation, et il faut calculer les corrélations multiples. Les résultats du calcul sont les suivants :

Hypoth.	R^2	$R^2_{[r]}$	n	k	r
1.	.451	.417	24	6	2
2.	.417	.400	24	4	1
3.	.400	.386	24	3	1
4.	.386	.226	24	2	1

En se servant du procédé conventionnel, on obtiendrait (cf. p.ex. *Kemphorne*, 1952) :

Hypoth.	$F = \frac{R^2 - R^2_{[r]} n - k - 1}{1 - R^2} r$	d.l.	Résultat
1.	.526	17 2	non significatif
2.	.554	19 1	non significatif
3.	.467	20 1	non significatif
4.	5.472	21 1	significatif 5 %

on arrive finalement à la conclusion qu'il faut utiliser les variables A/8 et A/4.

En se servant du procédé suggéré dans cette thèse, on obtient, du moment qu'on doit prédire x_0 (c'est-à-dire le succès individuel de chaque apprenti), les résultats suivants :

Hypoth.	$\left(\frac{R^2 - R_{[r]}^2}{1 - R_{[r]}^2}\right)^{\frac{1}{2}}$	$\left(1 - \frac{\varepsilon_{k-rn}}{\varepsilon_{kn}}\right)^{\frac{1}{2}}$	N	r	Résultat
1.	.241	.120	20	2	non significatif
2.	.168	.237	21	1	non significatif
3.	.151	.230	22	1	non significatif
4.	.455	.226	23	1	non significatif

Ce critère suggère donc l'utilisation d'une variable seulement. Dans ce cas, on devrait utiliser A/9 au lieu de A/8, car la corrélation entre A/9 et l'estimateur du succès est plus forte que lorsqu'on prend A/8. (A/8 n'a été choisie comme première variable que parce que, si tant est que l'on a besoin d'une combinaison de deux variables, celle des variables A/8 et A/4 donnerait des prédictions meilleures que l'utilisation d'une seule variable.)

L'estimation non biaisée de l'espérance mathématique de la longueur de l'intervalle de confiance relatif à un estimateur du succès, est donnée par :

$$\hat{E}(l_v) = [\varepsilon_{1\ 24} I_{00} (1 - R^2)]^{\frac{1}{2}} \alpha_{1\ 24} = 61.44.$$

7^o Appendice

7.1. Le calcul des tables

Les percentiles de t ne se trouvant pas dans des tables existantes, ont été calculés en utilisant les approximations dues à *Goldberg* et *Levine* (1946).

Les fonctions gamma ont été calculées logarithmiquement d'après les tables de *Davis* (1933).

La série hypergéométrique $F\left(\frac{k}{2}, \frac{n}{2}; \frac{n-1}{2}; -\frac{n}{n+1}\right)$ a été calculée :

pour $k = 1, 3$ par intégration numérique de

$$\begin{aligned} & \int_0^1 t^{k-1} (1-t)^{\frac{n-k-3}{2}} \left(1 - \frac{n}{n+1} t\right)^{-\frac{n}{2}} dt \\ &= F\left(\frac{k}{2}, \frac{n}{2}; \frac{n-1}{2}; \frac{n}{n+1}\right) B\left(\frac{k}{2}, \frac{n-k-1}{2}\right), \end{aligned} \quad (69)$$

en appliquant la règle de *Simpson*;

pour $k = 2, 4, 6, 8, 10, 14, 20$,

en partageant la série lentement convergente en deux séries rapidement convergentes, utilisant

$$\begin{aligned}
 F\left(\frac{k}{2}, \frac{n}{2}; \frac{n-1}{2}; \frac{n}{n+1}\right) \\
 = \frac{\Gamma\left(\frac{n-1}{2}\right) \Gamma\left(-\frac{k+1}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right) \Gamma\left(-\frac{1}{2}\right)} F\left(\frac{k}{2}, \frac{n}{2}; \frac{k+3}{2}; \frac{1}{n+1}\right) \\
 + \frac{\Gamma\left(\frac{n-1}{2}\right) \Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{k}{2}\right)} (n+1)^{\frac{k+1}{2}} F\left(\frac{n-k-1}{2}, -\frac{1}{2}; -\frac{k-1}{2}; \frac{1}{n+1}\right)
 \end{aligned} \tag{70}$$

(pour k pair, $k \geq 2$).

Quelques-unes des courbes représentant les limites de confiance inférieures relatives aux corrélations multiples, ont été copiées de *Ezekiel* (1929). Les courbes pour $n = 150, 200, 400$, et les courbes, pour 2, 4, et 6 variables indépendantes et n arbitraire, ont été calculées en utilisant la répartition non centrée de χ^2 comme approximation de la répartition du coefficient de corrélation multiple R dans les grands échantillons. (*Fisher*, 1928.)

Des quatre chiffres significatifs indiqués pour δ_{1n} et δ_{3n} , il n'en est parfois que les deux premiers qui soient précis; tous les autres chiffres donnés dans les tables sont précis.

n	k	1	2	3	4	6	8	10	14	20
20		2.056	2.118	2.185	2.259	2.433	2.654	2.949	4.044	
30		2.036	2.074	2.114	2.157	2.251	2.356	2.482	2.803	3.659
40		2.027	2.054	2.083	2.113	2.177	2.247	2.325	2.507	2.885
50		2.021	2.043	2.065	2.088	2.137	2.189	2.245	2.371	2.609
75		2.014	2.028	2.042	2.057	2.087	2.119	2.152	2.223	2.344
100		2.010	2.021	2.031	2.042	2.064	2.086	2.111	2.160	2.240
125		2.008	2.016	2.025	2.033	2.051	2.068	2.086	2.124	2.185
150		2.007	2.014	2.020	2.027	2.042	2.056	2.071	2.101	2.150
200		2.005	2.010	2.015	2.020	2.031	2.042	2.053	2.075	2.109

7.2. Tables de α_{kn} .

n	k	1	2	3	4	6	8	10	14	20
20		5.150	5.500	5.899	6.358	7.537	9.248	11.94	27.76	
30		4.644	4.834	5.040	5.262	5.770	6.387	7.147	9.385	17.62
40		4.421	4.549	4.684	4.827	5.142	5.501	5.913	6.956	9.449
50		4.295	4.391	4.492	4.598	4.824	5.073	5.349	6.005	7.355
75		4.135	4.194	4.255	4.318	4.451	4.590	4.739	5.067	5.657
100		4.058	4.102	4.145	4.190	4.282	4.379	4.480	4.698	5.065
125		4.014	4.047	4.081	4.116	4.187	4.261	4.337	4.498	4.764
150		3.985	4.012	4.039	4.068	4.125	4.186	4.246	4.374	4.583
200		3.948	3.968	3.988	4.010	4.052	4.094	4.141	4.229	4.372
400		3.894	3.904	3.914	3.924	3.944	3.964	3.985	4.026	4.091

7.3. Tables de ε_{kn} .

n	k	1	2	3	4	6	8	10	14	20
20		(.3957)	.7490	(1.096)	1.521	2.522	4.033	6.428	20.64	
30		(.2374)	.4396	(.6099)	.8409	1.280	1.844	2.547	4.613	12.24
40		(.1697)	.3105	(.4226)	.5795	.8537	1.187	1.575	2.555	4.904
50		(.1313)	.2400	(.3199)	.4420	.6399	.8745	1.138	1.761	3.047
75		(.08476)	.1529	(.2043)	.2773	.3929	.5263	.6704	.9862	1.557
100		(.06242)	.1122	(.1493)	.2020	.2834	.3761	.4748	.6858	1.044
125		(.04943)	.08863	(.1177)	.1589	.2216	.2926	.3675	.5250	.7851
150		(.04110)	.07323	(.1002)	.1309	.1819	.2394	.2997	.4252	.6289
200		(.03038)	.05434	(.07196)	.09680	.1339	.1756	.2189	.3081	.4498

7.4. Tables de δ_{kn} .

7.5. Graphique des limites de confiance à 95% inférieures relatives au coefficient de corrélation multiple.

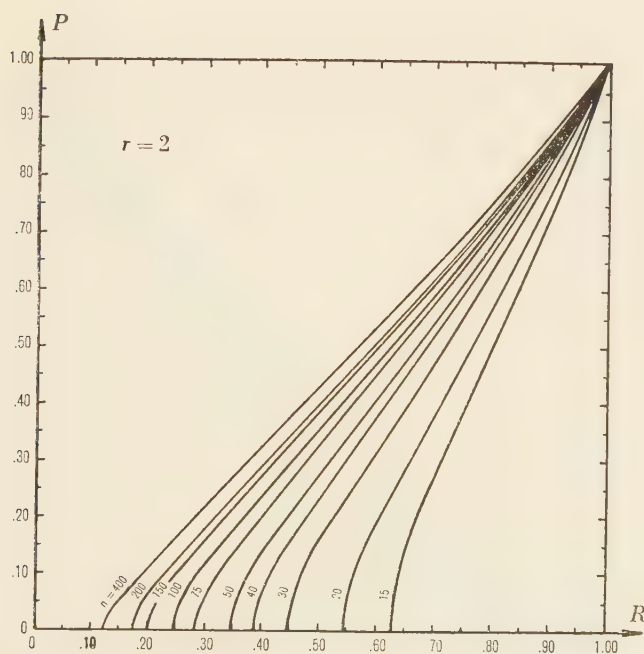
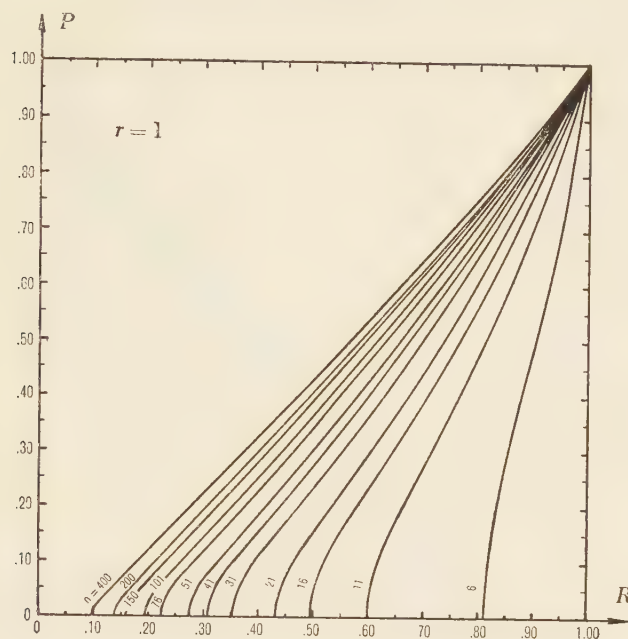
P coefficient de corrélation multiple dans l'ensemble de base.

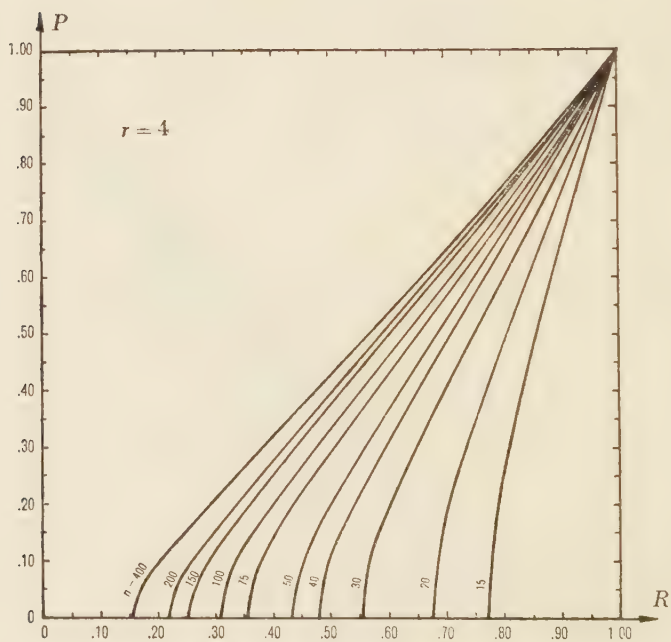
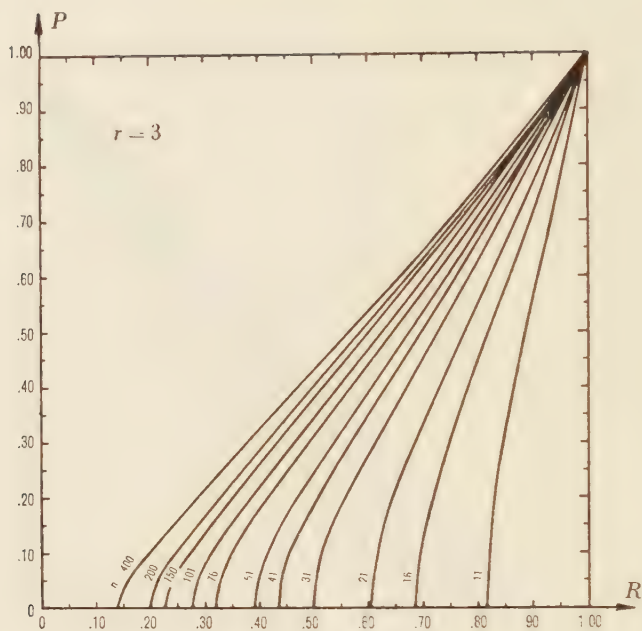
R coefficient de corrélation multiple dans l'échantillon.

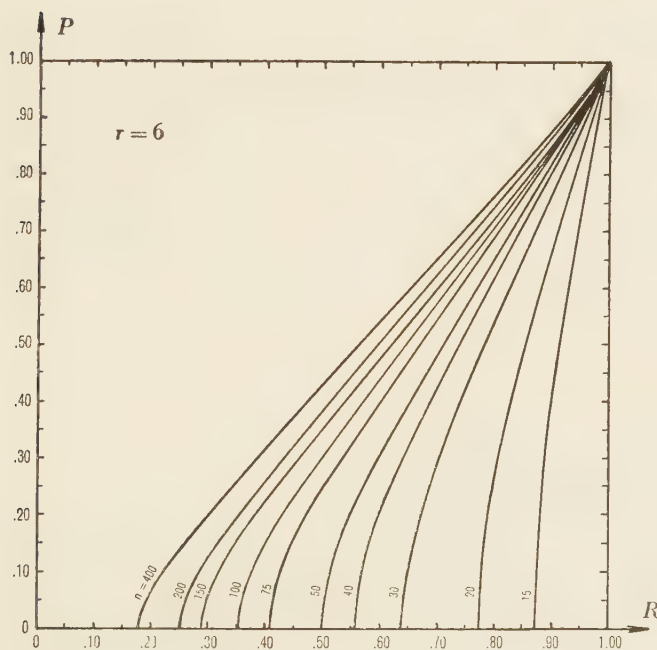
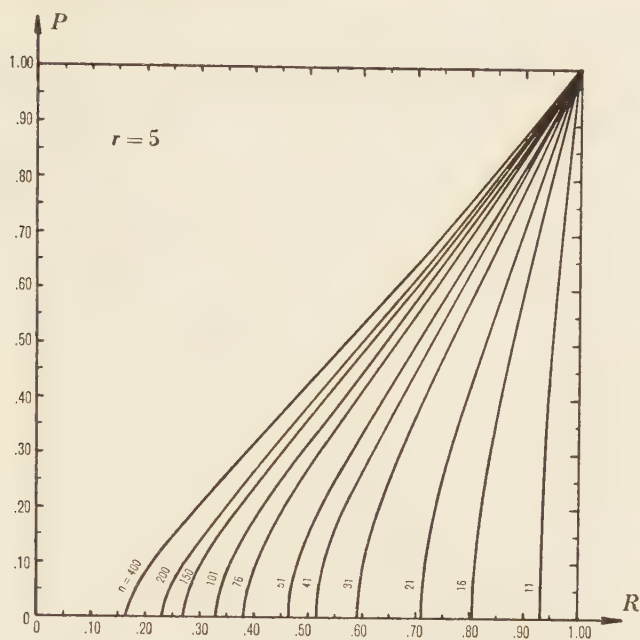
r nombre de variables indépendantes.

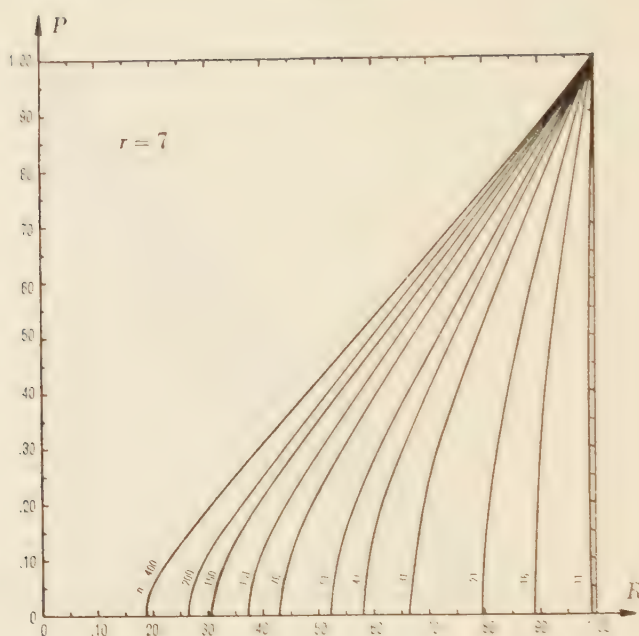
n effectif de l'échantillon.

Les courbes pour $r = 1, 3, 5$ et 7 ont été recopiées de *M. Ezekiel* (1929): *Methods of correlation analysis*, New York, John Wiley and Sons.









8° Remerciements

Nous remercions le Dr P. *Armsen* de nous avoir initié à la statistique mathématique. Nous sommes redevable au Dr S. *Biesheuvel* et au « South African Council for Scientific and Industrial Research » de l'autorisation de publier le présent travail. Nous tenons à exprimer notre gratitude également au professeur A. *Linder* d'avoir accepté la direction de notre thèse: ses précieuses remarques ont été d'un grand secours lors de la rédaction définitive de notre travail. Nos remerciements vont en outre au Dr H. S. *Steyn* d'avoir bien voulu lire des rédactions antérieures.

Nous sommes enfin obligé à M. R. *Lang* ainsi qu'à M^{lle} R. *Linder* de s'être chargés de mettre au point le manuscrit définitif.

Summary

A criterion is proposed for deciding whether r out of k given variables should be included in a regression analysis or not.

The criterion is restricted to the case where the k predictor variates and the variable to be predicted have a $(k - 1)$ -dimensional normal distribution. It is based on the expected value of the length of the confidence interval for the variable to be predicted or for its mean. The idea is to include the variables if this expected value becomes smaller and that is not always the case.

In sections 1 and 2 the idea of the criterion is introduced; in 3, the notation is given and in 4, the mentioned expected values are obtained. The distribution of a statistic which is needed for a test is found in section 5. The last sections contain instructions for applications, an example, tables of some constants and graphs of lower confidence limits for multiple R .

The computations which are necessary for the proposed criterion are only slightly heavier than those of the conventional F -test of significance of regression coefficients. The distribution obtained in section 5 is the power function of the F -test for the case of multivariate normal variables. The integral of section 5.3. may be of value in other problems of multivariate analysis. The proposed criterion is much stricter than the conventional one, particularly if the mean has to be predicted. Much less variables will be used.

Bibliographie

- Bartlett, M. S. (1933): On the theory of statistical regression. Proc. Roy. Soc. Edinburgh, 53, 260-283.
- Cramér, H. (1946): Mathematical methods of statistics. Princeton, University Press.
- Crow, E. L. (1955): Generality of confidence intervals for a regression function. J. Am. Stat. Ass. 50, 850-853.
- Davis, H. T. (1933): Tables of the higher mathematical functions. Vol. I. Bloomington, Principia Press.
- Ezekiel, M. (1929): Methods of correlation analysis. New York, John Wiley.
- Fisher, R. A. (1915): Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 10, 507-521.
- (1922): The goodness of fit of regression formulae and the distribution of regression coefficients. J. Roy. Stat. Soc., 85, 597-612.
- (1924): The distribution of the partial correlation coefficient. Metron, 3, 329-332.
- (1928): The general sampling distribution of the multiple correlation coefficient. Proc. Roy. Soc., A, 121, 654-673.
- Gauss, C. F. (1809): Theoria motus corporum coelestium. C. F. Gauss' Werke, Bd. Göttingen, 1873, Königl. Ges. d. Wissenschaften.
- Goldberg, H., Levine, H. (1946): Approximate formulas for the percentage points and normalization of t and χ^2 . Annals Math. Stat., 17, 216-225.
- Guttman, L. (1954): A new approach to factor analysis: The radex. Mathematical thinking in the social sciences. (Lazarsfeld Ed.) Glencoe, Free Press. p. 307.
- Hotelling, H. (1931): The generalization of Students ratio. Annals Math. Stat., 2, 359-378.
- Hsu, P. L. (1938): Notes on Hotellings generalized T . Annals Math. Stat., 9, 231-243.
- Johnson, P. O. (1949): Statistical methods in research. New York, Prentice Hall.
- Kemphorne, O. (1952): The design and analysis of experiments. New York, John Wiley.
- Legendre (1806): Nouvelles méthodes pour la détermination des orbites des comètes. Paris.
- Magnus, W., Oberhettinger, F. (1943): Formeln und Sätze für die speziellen Funktionen der mathematischen Physik. Berlin, Springer.

- Stouffer, E. B.* (1932): Expression for a determinant in terms of five minors. *Am. Math. Monthly*, 39, 165-166.
- Verwey, F. A.* (1957): Prediction of work performance of artisan apprentices during training by means of psychological tests. Unpublished Ph. D. thesis, Univ. of the Witwatersrand, Johannesburg.
- Wilks, S. S.* (1932): Certain generalizations in the analysis of variance. *Biometrika*, 24, 471-494.
- Wishart, J.* (1928): The generalized product moment distribution in samples from a normal multivariate distribution. *Biometrika*, 20A, 32-52.
- Yule, G. U.* (1907): On the theory of correlation for any number of variables treated by a new system of notation. *Proc. Roy. Soc. A.*, 79, 182-193.

